Procedures and Problems in Korean-Chinese-Japanese Wordnet with Shared Semantic Hierarchy

Key-Sun Choi and Hee-Sook Bae

KORTERM, KAIST
373-1 Guseong-dong, Yuseong-gu, Daejeon, Republic of Korea
Email: {kschoi,elle}@world.kaist.ac.kr

Abstract. This paper introduces a Korean-Chinese-Japanese wordnet for nouns, verbs and adjectives. This wordnet is constructed based on a hierarchy of shared semantic categories originated from NTT Goidaikei (Hierarchical Lexical System). The Korean wordnet has been constructed by mapping a semantic category to each Korean word sense in a way that maps the same semantic hierarchy to the meanings of nouns, verbs, and adjectives. The meaning of each verb searched in the corpus is compared with its Japanese equivalent. The Chinese wordnet has been also constructed based on the same semantic hierarchy in comparison with the Korean wordnet. In terms of the argument structure, there is a semantic correspondence between Korean, Japanese and Chinese verbs.

1 Introduction

A Korean-Chinese-Japanese wordnet named CoreNet has been developed using a shared semantic hierarchy since 1994. This semantic hierarchy is originated in NTT Goidaikei[1], which consists of 2,710 hierarchical semantic categories. For the purpose of this paper, the term "wordnet" refers to a network of words, the term "concept" to the semantic category, and the term "sense" to the different meaning of word. In CoreNet, a total of 2,954 concepts are specified. An increase in the number of concepts specified in CoreNet is attributable to the necessity for reflecting the concepts found only in the Korean language. On the one hand, the same semantic hierarchy applied to both nouns and predicates in CoreNet, while different concept systems are applied to nouns and predicates in NTT Goidaikei.

Mapping the same semantic hierarchy to both nouns and predicates results in some advantages: first, there are pattern similarities between nouns and predicates, especially in Chinese-derived words (that is N in the following example). For example, "N-hada and "N+suru" are the Korean and Japanese version of a basic pattern "do + N" in English; second, the language generation based on a conceptual structure takes freer phrase patterns regardless of either the noun or verb. This computational work has been accompanied by heuristics and trial-and-errors as well as semi-automatic approaches. Several linguistic resources have been used for building CoreNet. Among them, [2] and [3] have been primarily used as a basis for the meanings of Korean words. Most of the Chinese vocabulary is based on [5].

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 91–96.

2 Principles

CoreNet has been constructed according to the following principles: multiple mapping between the word sense and the concept, corpus-based, multilingualism, and application of a single concept system.

2.1 Mapping between Word Sense and Concept

The purpose of CoreNet is mainly to resolve semantic ambiguities using the following two functionalities. Firstly, every possible meaning of a word in the dictionary [3] is mapped to one or more concepts. For example, each meaning of the word "*school*" is mapped into three concepts; PLACE, ORGANIZATION, and BUILDING. In the second place, a syntactic-semantic structure is mapped to the predicate-argument structure. For example, a Korean verb "*gada*" has a set of 17 senses in the dictionary [3]; these word senses are mapped into the concepts such as GOING, LEARNING, SERVICE, DELIVERY, PROGRESS, CONTINUATION, ENTHUSIASM, SWEEP, and so on. This set of predicate concepts is identical to nouns'. On the other hand, each predicate has its unique argument structure. For example, "*gada*" is mapped into seven concepts (e.g., GOING, LEARNING) whose argument structures are different. Each argument is represented by the set of possible concept filler (e.g., [HUMAN]) and syntactic role (e.g., subject, dative, and object) while its Japanese equivalents (e.g., "*iku*") are addressed by the followings:

1. GOING([HUMAN,MAMMAL,VEHICLE]=subject), "*iku*"
2. LEARNING([HUMAN]=subject,[TEACHER]=dative), "*iku"*
3. DELIVERY([INFORMATION]=subject,[HUMAN]=dative), "*tutawaru*"
4. PROGRESS([TIME]=subject), "*sugiru*"
5. CONTINUATION([RELATION]=subject,[YEAR]=object), "*tuduku*"
6. ENTHUSIASM([GAZE]=subject,[GIRL]=dative), "*iku*"
7. SWEEP([EMOTION]=subj), "*kieru*"

2.2 Corpus-based usage

A set of vocabularies and their meanings are extracted from KAIST corpus [2]. The following shows what the argument structure of "*gada*" described in the section 2.1 is like when extracted from the corpus: GOING ([*horse*/MAMMAL, *bus*/VEHICLE]=subject)

Horse and *bus* are the terms extracted from the corpus while MAMMAL and VEHICLE are the concept names respectively mapped to the words *horse* and *bus*. This results in more specified categorization for the meaning of words than in dictionaries.

2.3 Multilingualism

All concepts are aligned with three languages: Japanese, Korean and Chinese. Among these three languages, all words that are nouns or predicates are categorized into a single concept hierarchy. Based on the meanings of words as well as concepts, verbs among three languages are also linked each other. The following is part of a list of concepts for the Chinese verb [qù]. Note that the *italicized* words are Korean equivalents. A sample list is shown in Figure 1.

1. GOING - *gada*
2. DELIVERY – *bonaeda*
3. EXCLUSION - *eobsaeda*

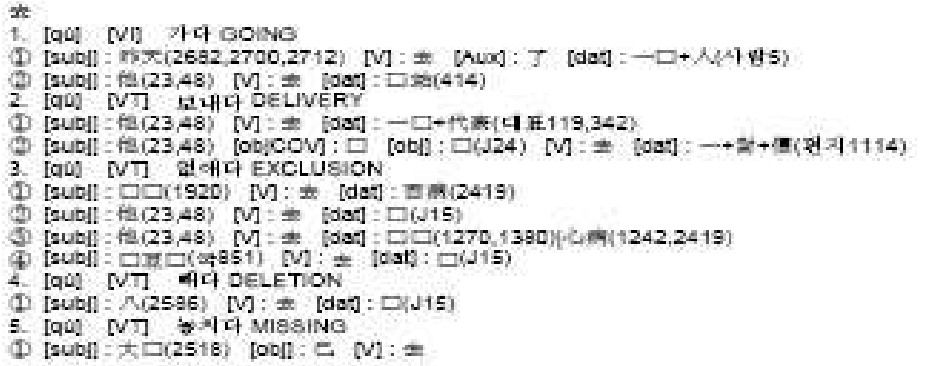

Fig. 1. An Entry in Chinese-Korean Verb CoreNet

2.4 Single Concept System

In general, concept systems and word nets are constructed for nouns. In CoreNet, however, a single concept system is shared by nouns, verbs, and adjectives. To this respect updates are continuously made for sharing of single concept system among three languages.

3 Procedures

3.1 Selection of Word Entry

A set of basic words is selected from the frequency-based vocabulary list of corpora compared with an existing set of basic Korean words. About 50,000 general vocabularies are selected for CoreNet word entries.

3.2 Bootstrapping for Initial Semantic Category Assignment

Using a Japanese-Korean electronic dictionary, we translated all Japanese words in the NTT Goidaikei into their Korean equivalents based on word meanings. Manual correction by experts of the results of automatic translation is followed for erroneous assignments between the two languages. This process also poses many problems. The most difficult problem issues from the difference in concept division systems. In Japanese, for example, concepts like GOING or SORTING have more subordinates than in Korean language, and vice versa for ROOT. In addition, FURNITURE has subordinate concepts like DESK, CHAIR, and FIREPLACE,

while in Korean, FIREPLACE is dealt with as part of KITCHEN. These problems arise from the difference in the way of thinking and culture. Then, we assign a semantic category by matching Korean words with their equivalent list for the semantic category in the NTT Goidaikei. No equivalent can be found in the translated word list and some errors can be found in a translation version. In the former case, a genus term for the word is extracted from descriptive statements of a machine-readable dictionary. In the latter case, manual correction is performed by experts.

3.3 Semantic Category Assignment Based on Word Sense Definitions [4]

Assuming that meanings falling under a concept are defined by similar words in the dictionary, we collected the definitions of the word senses that were mapped into one concept incorporating them into the concept's definition. This resulted in the creation of a chunk of definitions per concept. That is, the definition of a concept is indirectly represented by the chunk of definition of word senses that has already been assigned to the concept. For a given new word sense, its appropriate concept assignment is to be solved by how much the definition of the word sense is similar with the definition of concept. Assignment of proper concepts to the word sense can be viewed as retrieving a relevant definition chunk (of concept) for the given word sense. Each concept's definition is incrementally upgraded whenever the definition for a new word sense is assigned to the concept.

Our structured version of the Korean dictionary [3] includes such lexical relation information as synonyms, abbreviations, antonyms, *etc*. It is reasonable that the two senses linked by this lexical relation information (except for antonyms) fall under the same concept.

3.4 Manual Correction

The process of resolving the meaning of a word (i.e. word sense disambiguation) was manually performed in order to assign proper semantic categories to every possible meaning of a word, as well as translation errors were removed. The same manual correction was independently performed by two researchers. After comparative review over the results, only identically mapped sets were selected as final semantic categories with the purpose of ensuring highest accuracy. In the final stage, a third party examined different parts of the results to choose the proper ones. Despite this manual correction, it remains still some embarrassing cases. For example, 출입(出入) is a word having a concept combined with two concepts GO OUT and ENTER. In this case, we selected the concept of superior node when the latter contains all of concept elements as following: 出入 [GO OUT-ENTER, 2183].

4 Considerations

This section describes what we had to consider and decide about the underspecified sense, multiple concept mapping, verbal noun, and concept splitting.

4.1 Underspecified Sense and Multiple Concept Mapping

A word is mapped into several concepts that comprise respective meanings of the word. For example, *school* is an "institution for the instruction of students". The word *school* is mapped

into three concepts such as LOCATION, ORGANIZATION, and FACILITY. Unless the meanings of a word are fully specified in the mother dictionary [3], however, one meaning of the word must be mapped into several concepts. The word *school* is a good example of underspecified meanings.

4.2 Verbal Noun

A verb is assigned to concepts after it is transformed to a noun. For example, "*write*" is transformed to its noun form "*writing*" that is mapped into a concept WRITING falling under EVENT. An adjective "*be wise*" is transformed to "*wisdom*" that is mapped into PROPERTY under CAPABILITY, which falls under ATTRIBUTE. Consider an adjective "*be wide*". A sense is mapped respectively to POSITIVE PERSONALITY, EXTENT/LIMITS, and WIDTH (under the concept UNIT OF QUANTITY).

4.3 Concept Splitting

Every time inconsistency among nodes of concepts is discovered, a node may be added. For example, BODY has three sub-concepts in the NTT concept system: ARM, LEG, and HEAD. But, a word "*back*" cannot be assigned to any sub-concepts. At least, OTHERBODY should be added to the fourth sub-concepts under BODY. In the course of constructing verbs and adjectives wordnets, the concept splitting was performed by reclassifying the word senses. For example, ARRIVAL is subdivided into SITUATION ARRIVAL, TIME ARRIVAL, EXTENT ARRIVAL, and POSITION ARRIVAL.

5 Example

Figure 2 shows a screenshot of the Korean-Japanese noun wordnet. The screen has four windows. The upper left side of the window shows a correspondence between Japanese and Korean words and concept numbers. The lower left side of the window contains word senses and definitions in the dictionary [3]. The upper right side of the window shows all words under a concept QUANTITY numbered 2588. The lower right side of the window shows a part of the list of concept hierarchy.

6 Conclusion

CoreNet has been constructed in combination with its necessary corpora and lexical database. To begin with, the keynote system of the NTT Goidaikei [1] was used, which was followed by the development of a Korean version of noun systems. Despite the different semantic categories applied to predicates in the NTT Goidaikei, we have aggressively applied the same semantic categories to predicate systems in CoreNet. Further, what differs between CoreNet and NTT Goidaikei is that CoreNet features mapping between word senses (not just words) and concepts. Multilinguality is another feature of CoreNet designed to deliver a single concept system for different languages.

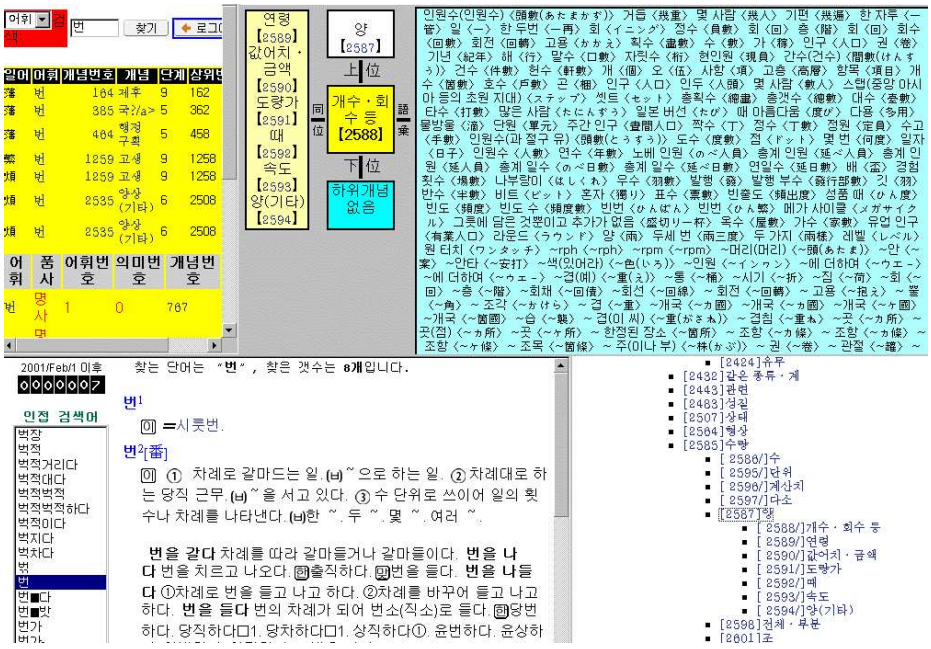

Fig. 2. A Screenshot of Korean-Japanese Noun Wordnet

References

1. Ikehara, S. *et al*: The Semantic System, volume 1 of Goidaikei – A Japanese Lexicon, Iwanami Shoten, Tokyo (1997).
2. KAIST Corpus, `http://morph.kaist.ac.kr/kcp/`, (in Korean), 1999–2003.
3. Hangeul Society, ed.: Urimal Korean Unabridged Dictionary, Eomungag (1997).
4. Lee, J.-H. *et al*.: Semi-Automatic Construction of Korean Noun Thesaurus by Utilizing Monolingual MRD and an Existing Thesaurus, Proceedings of the 16th PACLIC, Jeju (2002).
5. Yu, S.: Modern Chinese Grammar Information Dictionary, Peking University Press (1999).